Automatic Analysis of Hydrogen/Deuterium Exchange Mass Spectra of Peptides and Proteins Using Calculations of Isotopic Distributions

Magnus Palmblad, Jos Buijs,* and Per Håkansson
Division of Ion Physics, Ångström Laboratory, Uppsala University, Uppsala, Sweden

High mass-resolving power has been shown to be useful for studying the conformational dynamics of proteins by hydrogen/deuterium (H/D) exchange. A computer algorithm was developed that automatically identifies peptides and their extent of deuterium incorporation from H/D exchange mass spectra of enzymatic digests or fragment ions produced by collisionally induced dissociation (CID) or electron capture dissociation (ECD). The computer algorithm compares measured and calculated isotopic distributions and uses a fast calculation of isotopic distributions using the fast Fourier transform (FFT). The algorithm facilitates rapid and automated analysis of H/D exchange mass spectra suitable for high-throughput approaches to the study of peptide and protein structures. The algorithm also makes the identification independent on comparisons with undeuterated control samples. The applicability of the algorithm was demonstrated on simulated isotopic distributions as well as on experimental data, such as Fourier transform ion cyclotron resonance (FTICR) mass spectra of myoglobin peptic digests, and CID and ECD spectra of substance P. (J Am Soc Mass Spectrom 2001, 12, 1153–1162) © 2001 American Society for Mass Spectrometry

Mass spectrometry is rapidly advancing as a major tool in proteomics and structural genomics because of its sensitivity, speed, and simplicity in identifying protein sequences, post-translational modifications, and protein structures [1–3]. Especially, electrospray ionization (ESI) combined with FTICR mass spectrometry has become a powerful and efficient tool for probing protein structure and function [4, 5]. Due to the high mass-resolving power and mass accuracy routinely achieved with high-field superconducting magnets [6], fragment ions covering parts of the protein sequence are readily assigned even out of a complex mixture of peptides [2, 7]. Furthermore, analysis of isotopic patterns assists in charge state and accurate mass determination [8, 9] and can reveal structural aspects, such as the oxidation state of metalloproteins [10], or information on the atomic composition of a protein [11].

High mass-resolving power mass spectrometry is also very useful for studying the conformational dynamics of proteins by monitoring the H/D exchange process. The analytical power in probing the H/D exchange process is that in a structured protein, amide hydrogen exchange is dramatically slowed down compared to exchange rates in random coil-like peptides [12–14]. The sensitivity of the H/D exchange process of amide hydrogens towards the structural flexibility of protein structures provides an experimental window to study protein folding [15, 16], structure [17–19], stability [20, 21], and protein–ligand interactions [22].

To take full advantage of mass spectrometry in structural studies of proteins, it is useful to take the isotopic patterns produced by biomolecules into account [23]. These patterns, or isotopic distributions, depend on the relative abundances of different isotopes of the constituting elements and the elemental composition of the molecular species. For protein samples containing natural abundances of isotopes, a method to estimate the isotopic distributions has been developed by Senko et al. [9], assuming that isotopic peaks can be resolved and that only the approximate molecular mass is given. This method uses an average amino acid, averagine, derived from a large number of proteins in a sequence database, and calculates the expected isotopic distribution of a poly-averagine of molecular mass close to the measured molecular mass. The method has been successfully implemented in a computer algorithm to automatically identify isotopic clusters in complex mass spectra [24]. In most H/D exchange studies, the measured masses are centroid values which take into ac-

Published online August 23, 2001

Address reprint requests to M. Palmblad, Division of Ion Physics, Ångström Laboratory, Uppsala University, Box 534, SE-751 21 Uppsala, Sweden. E-mail: magnus.palmblad@angstrom.uu.se

* Current address: Biacore, Rapsgatan 7, SE-754 50 Uppsala, Sweden.

1044-0305/01/$20.00
PII S1044-0305(01)00301-4

Received February 28, 2001
Revised July 6, 2001
Accepted July 9, 2001

count the asymmetry of the peak arising from the natural isotopic distribution [22, 23, 25]. For partially deuterated proteins, the measured isotopic distributions can be approximated by normal distributions, and such an approximation has for example been used to evaluate the heterogeneity of conformational states of proteins [16, 23].

Many H/D exchange studies are aimed at getting a high structural specificity, which can be obtained if the proteins are fragmented after the H/D exchange step. This is for example achieved by enzymatic digestion [13, 21, 25] or by physical fragmentation techniques [26–28], such as CID [29–31] and ECD [32, 33]. If the sequence of the protein is known, this sequence information can be used to generate a short list of candidates for the identity of an isotopic cluster, given the particular fragmentation method used. To rank the candidates, the pattern of isotopic peaks can be used together with the measured accurate masses.

Upon H/D exchange, the ratio of hydrogen to deuterium of labile hydrogens in the molecule increases over time. Thus, the isotopic distribution has a known component, derived from a molecular formula and known isotopic abundances, and an unknown component, dependent on the exchange rates of the exchanging hydrogens, which in turns depends on the structure and dynamics of the peptide or protein. When deuterium is incorporated into the molecule, the isotopic distribution shifts to higher masses, but as it does so, the distribution also changes. Particularly for smaller species this alteration in the distribution is readily measured by mass spectrometry. The whole distribution contains not only information on the average extent of deuterium incorporation but also on the variation in the extent of deuteration [13, 23]. Measurements of the whole isotopic distribution over time can be used to calculate exchange rates of the labile hydrogens, for instance by adapting the maximum entropy method (MEM) used by Zhang et al. [34].

It will be shown in this paper that accurate mass determination (to within a few ppm) and comparison of experimental and calculated isotopic distributions are together often necessary and also sufficient to identify all components in H/D exchange mass spectra and evaluate their extent of deuterium incorporation. A computer algorithm performing such analyses of H/D exchange mass spectra will be described and demonstrated on peptic digests of myoglobin, and CID and ECD spectra of substance P, making identification of peptic and CID/ECD fragments fully automated and significantly less dependent on undeuterated controls. As additional examples of the universality of the approach, the algorithm will be tested on simulated spectra of the cysteine-rich protein metallothionein and an enzymatic fragment of cytochrome *c* containing the heme group, as well as on a simulation of peptic digests of larger proteins. Although the present paper is focused on data processing, H/D exchange experiments are set up in a flow-through system, without any chromatographic separation step, which can easily be automated. It should be noted that the approach presented in this paper is not dependent on the resolving power of FTICR mass spectrometry, but is applicable also for other types of mass analyzers, such as quadrupole ion traps and time-of-flight instruments, since information on the isotopic distributions is available also from these.

Materials and Methods

Sample Preparation

Myoglobin from sheep skeletal muscle and substance P were obtained from Sigma Chemical Co., (St. Louis, MO) and used without further purification. For H/D exchange experiments with myoglobin, 1 mM myoglobin was prepared in a 100 mM Hepes (Merck KGaA, Darmstadt, Germany) buffer in H_2O. Deuterium exchange was initiated by diluting the myoglobin solution 20 times in D_2O (>99.8%, Merck). This solution contains 50 μM myoglobin in 5 mM Hepes, pH 6.5 (no isotopic correction), and a D content of 95%. All above mentioned solutions were at 21.5 °C. After certain exchange times, aliquots of 20 μL were taken and myoglobin was digested with pepsin by adding 130 μL of a 6 μM pepsin solution dissolved in a 6% acetic acid/H_2O at 0 °C. The lowered pH (from 6.5 to 3.0) and temperature slows down the H/D exchange reaction with a factor of 8×10^3 [35]. After digestion for one min, 130 μL, 0 °C methanol was added and the mass spectrometric acquisition was started as described below. The preparation of deuterated substance P samples has been described elsewhere [36].

Mass Spectrometry

The sample was rapidly loaded in a syringe (250 μL, Hamilton Co., Reno, NV) and infused into an electrospray ion source (Analytica, Branford, CT) using a syringe pump (model Sage 361, Thermo Orion, Beverly, CA) at a flow rate of 2.0 μL/min. Ion formation was assisted by using nitrogen as nebulizing gas. The ion source was constantly purged with dry nitrogen gas at ambient temperature. The ion source was linked to the Analytica atmosphere–vacuum interface and a potential difference between the spraying needle and the inlet capillary of 4 kV was applied. All mass spectra were acquired using a Bruker Daltonics (Bruker, MA) 9.4 tesla FTICR mass spectrometer. A general description of the instrument and its performance characteristics have been published elsewhere [37].

Computer Analysis

Given a molecular formula, it is straightforward to calculate the isotopic distribution by convoluting the isotopic distributions of the composing elements. A rapid method of calculating the convolution $f*g$ of two

functions f and g is to use the convolution theorem in Fourier analysis

$$\mathcal{F}[f*g] = \mathcal{F}[f]\mathcal{F}[g] \quad (1)$$

where F is the Fourier transform. In the discrete case, f and g are represented by vectors, and in our case, these vectors are the isotopic distributions of the elements. For example, the isotopic distribution of carbon, here denoted $\mathbf{a}_C$, can be written as

$$\mathbf{a}_C = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0.98889\ 0.01111\ 0\ \ldots\ 0) \quad (2)$$

The index of a component encodes the nominal mass whereas the components are the relative abundances of the isotopes (here ^{12}C and ^{13}C). The convolution is calculated by componentwise multiplication of the Fourier transforms of the isotopic distributions of the elements, followed by inverse transformation of the product [38].

$$\mathbf{a}_{protein} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{a}_H)^{n_H}\mathcal{F}(\mathbf{a}_C)^{n_C}\mathcal{F}(\mathbf{a}_N)^{n_N}\mathcal{F}(\mathbf{a}_O)^{n_O}\mathcal{F}(\mathbf{a}_S)^{n_S}) \quad (3)$$

In 3, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the discrete Fourier transform and the inverse discrete Fourier transform respectively, $\mathbf{a}_H$, $\mathbf{a}_N$, $\mathbf{a}_O$, and $\mathbf{a}_S$ are the isotopic distributions of hydrogen, nitrogen, oxygen and sulphur, respectively, defined analogously to $\mathbf{a}_C$ in eq 2. n_H, n_C, n_N, n_O, and n_S are the number of atoms of the respective element in the molecule, combining to yield the isotopic distribution $\mathbf{a}_{protein}$ of the whole protein molecule. Discrete Fourier transforms are computed through a variant of the FFT algorithm [39] and the computations are accelerated by allowing oversampling of the isotopic distributions. The use of oversampling is based on the fact that vector size, i.e., the size of the vector used to store and calculate isotopic distributions, does not have to scale with the mass of the molecule but only with the width of the isotopic distribution. The number of times a distribution is folded over can easily be calculated, but it is not necessary since mass modulo vector size can be used to access both calculated and measured intensities. Calculated and measured masses are matched by comparing the integer and the monoisotopic mass of the candidates.

First, a program called DIG is run to perform in silico digestion or fragmentation of the protein under study. This program is identical to one that has previously been used for tryptic digests, except the definition of enzyme action is changed to pepsin digestion, CID, or ECD. The definition of the enzyme/fragmentation technique is read as input to the program. The output of DIG is all possible fragments, given the sequence and cleavage/fragmentation rules, post-translational modifications and constraints on fragment length, and missed cleavage sites, similar to MS-Digest [40]. The AUTOHD computer program (Figure 1) takes as input peak lists containing m/z values and intensities generated by the software supplied with the instrument and the times at which the H/D exchange reaction was quenched. The first step is to reduce the large amount of peaks to individual isotopic clusters [41]. Different charge states are treated separately by default as the charge may influence gas phase back-exchange reactions [42] or different charge states may belong to different conformational populations in the first place [43]. The second step is to identify each isotopic cluster, which is done by comparing calculated and experimental masses and relative intensities, assuming some distribution of isotopic ratios of the exchanging hydrogens. The computer program generates a list of candidates which possibly match the fragment, given the measured

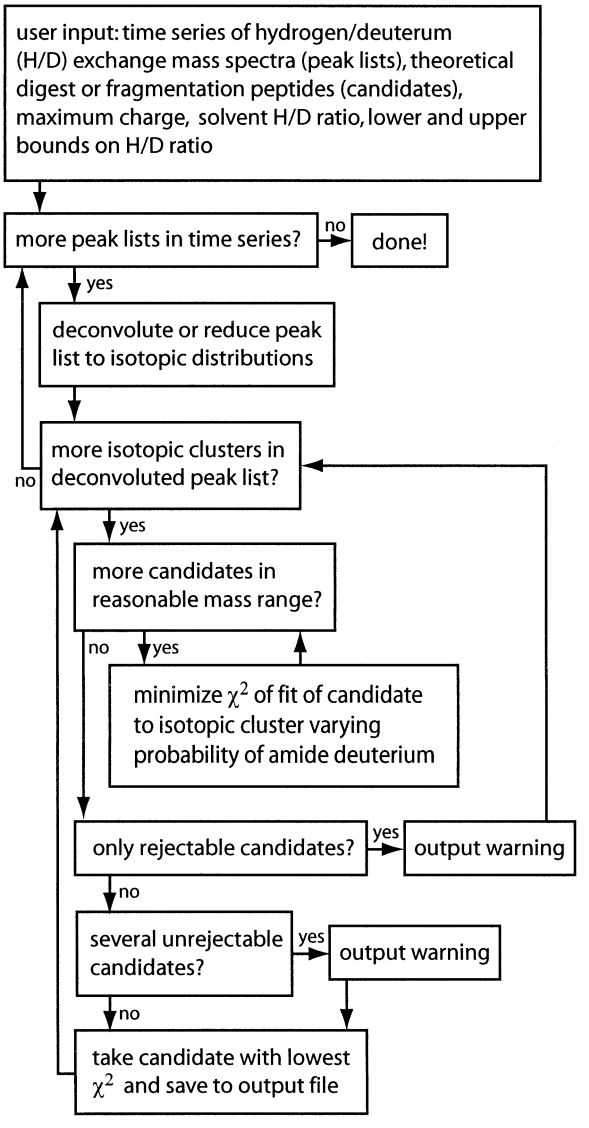

Figure 1. Flowchart of the AUTOHD program used in the automatic analysis of H/D exchange mass spectra. The program takes in a list of mass spectra, recognizes isotopic clusters, and tries to identify these using a list of candidates generated from the amino acid sequence and the enzymatic fragmentation or dissociation method used. The average amide H/D ratio is also calculated.

mass. A lower mass limit of the candidates is given by the mass at which 100% of the labile hydrogens would have to be deuterated for the candidate to reach the measured masses. The upper mass limit of the candidates is given by the measured masses. In the simplest case, the distribution of the exchanging hydrogens is assumed to be binomial

$$\mathbf{a}_{\text{labile}} = \text{Bin}(P(^2\text{H}), n_{\text{labile}}) \quad (4)$$

where $P(^2\text{H})$ is the probability of an exchanging (amide) hydrogen being a ^2H, $\mathbf{a}_{\text{labile}}$, and n_{labile} being defined as above. Eq 4 holds if all hydrogens have the same probability of being deuterated. This assumption is not true in general, in which case the binomial distribution serves as an approximation of the true distribution. However, it is straightforward to replace eq 4 with any other model of the isotopic distribution of labile hydrogens. Only the hydrogens exchanging on the time scale of the experiment are treated in eq 4. Hydrogens that are exchanging too quickly or too slowly to be measured are accounted for separately, i.e., they are either at equilibrium with the D content in the solution, $\mathbf{a}_{\text{solvent}}$, or follow the natural isotopic distribution, $\mathbf{a}_{\text{H}}$, /espectively. The theoretical distribution

$$\mathbf{a}_{\text{protein}} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{a}_{\text{H}})^{n_{\text{H}}}\mathcal{F}(\mathbf{a}_{\text{C}})^{n_{\text{C}}}\mathcal{F}(\mathbf{a}_{\text{N}})^{n_{\text{N}}}\mathcal{F}(\mathbf{a}_{\text{O}})^{n_{\text{O}}}\mathcal{F}(\mathbf{a}_{\text{S}})^{n_{\text{S}}}\mathcal{F}(\mathbf{a}_{\text{solvent}})^{n_{\text{solvent}}}\mathcal{F}(\mathbf{a}_{\text{labile}})) \quad (5)$$

is then fitted to experimental data by minimizing the χ^2-value (eq 6) using only one free parameter $P(^2\text{H})$, for different candidates from the generated list of possible fragments. In other words, the isotopic distribution of a fragment candidate given a certain $P(^2\text{H})$ is calculated by first transforming and multiplying the isotopic distributions of all elements and atoms not exchanging on the time scale of the experiment, then multiplying this vector, component by component, with the transform of the binomial amide hydrogen distribution and finally calculating the inverse transform of the product vector which gives the isotopic distribution. The product $\mathcal{F}(\mathbf{a}_{\text{H}})^{n_{\text{H}}}\ \mathcal{F}(\mathbf{a}_{\text{C}})^{n_{\text{C}}}\ \mathcal{F}(\mathbf{a}_{\text{N}})^{n_{\text{N}}}\ \mathcal{F}(\mathbf{a}_{\text{O}})^{n_{\text{O}}}\ \mathcal{F}(\mathbf{a}_{\text{S}})^{n_{\text{S}}}\ \mathcal{F}(\mathbf{a}_{\text{solvent}})^{n_{\text{solvent}}}$ is calculated once for every candidate and stored.

The calculated distribution is then fitted to experimental data using standard χ^2-statistics. It is assumed that the measurement errors in the intensity of the M isotopic peaks are independent and normal distributed, i.e., the χ^2-value is a sum of squared standard normal distributions N(0,1), which means that χ^2 is χ^2-distributed with $M - 1$ degrees of freedom since there is one linear constraint from the normalization of the intensities.

$$\chi^2 = \sum_{m=1\ldots M} \frac{(I_m^{\text{calculated}} - I_m^{\text{measured}})^2}{\sigma_m^2} \in \chi^2_{M-1} \quad (6)$$

The experimental data are fitted by minimizing this χ^2-value, varying the $P(^2\text{H})$. The χ^2-value is calculated from all measured isotopic peaks and corresponding calculated isotopic peaks, and also including peaks below the lowest measured mass and above the highest measured mass if the calculated intensities at these masses should have been detected. The minimum χ^2-value is reported along with the $P(^2\text{H})$ at this minimum, argmin(χ^2). The argmin(χ^2) can be interpreted as the average amide deuterium content in a peptide. The σ_2^m-value used in eq 6 has two terms, one from the experimental signal variance, which is approximately proportional to signal strength, and one from the random noise variance, which is constant, and added to the signal independently of the signal strength.

$$\sigma_m^2 = \sigma_{m,\text{signal}}^2 + \sigma_{\text{noise}}^2 \approx \text{c}I_m^{\text{measured}} + \sigma_{\text{noise}}^2 \quad (7)$$

The constant c was determined from peaks of similar mass-to-charge ratios and intensities in control experiments and σ_{noise}^2 is measured in the analyzed spectra.

To find the best candidate, the mass measurement error is incorporated in eq 6. This yields a total χ^2-value, and by calculating the corresponding quantile, the significance of rejecting the candidate and model for deuterium incorporation. Mass measurement errors are assumed to be normal distributed with an experimentally obtained variance [41]. If there are none or several unrejectable hypotheses, the program outputs a warning. Other constraints that can be used include lower and/or upper bounds on $P(^2\text{H})$ if these are known or can be assumed.

The one-dimensional minimization is performed by the Brent algorithm [44, 45], which converges in just a few iterations since χ^2 is a polynomial in $P(^2\text{H})$ and the coefficients of the higher powers in this polynomial are small, at least if $P(^2\text{H})$ is small. The Brent method iteratively fits a parabola to three points in an interval containing the minimum, then updates the three points by adding the minimum of the parabola and discarding the maximum point.

Implementation of Algorithms

In the algorithms described and used in this paper, the FFTW implementation [46] of the FFT algorithm was used for portability and speed. The peak reduction or deconvolution algorithm was written in OCaml [47] and compiled separately by the OCaml 3.00 compiler. This algorithm has been described elsewhere [41]. The rest of the software was written in ANSI C using the Brent minimization algorithm code [45] and compiled using GCC [48] 2.95.2 under both IRIX 6.3 and Windows 2000 using the Cygwin [49] driver.

The DIG program was run allowing up to twenty missed peptic cleavage sites with minimum fragment length of four residues. All b and y ions were generated for interpretation of CID spectra and all c and z ions for ECD spectra.

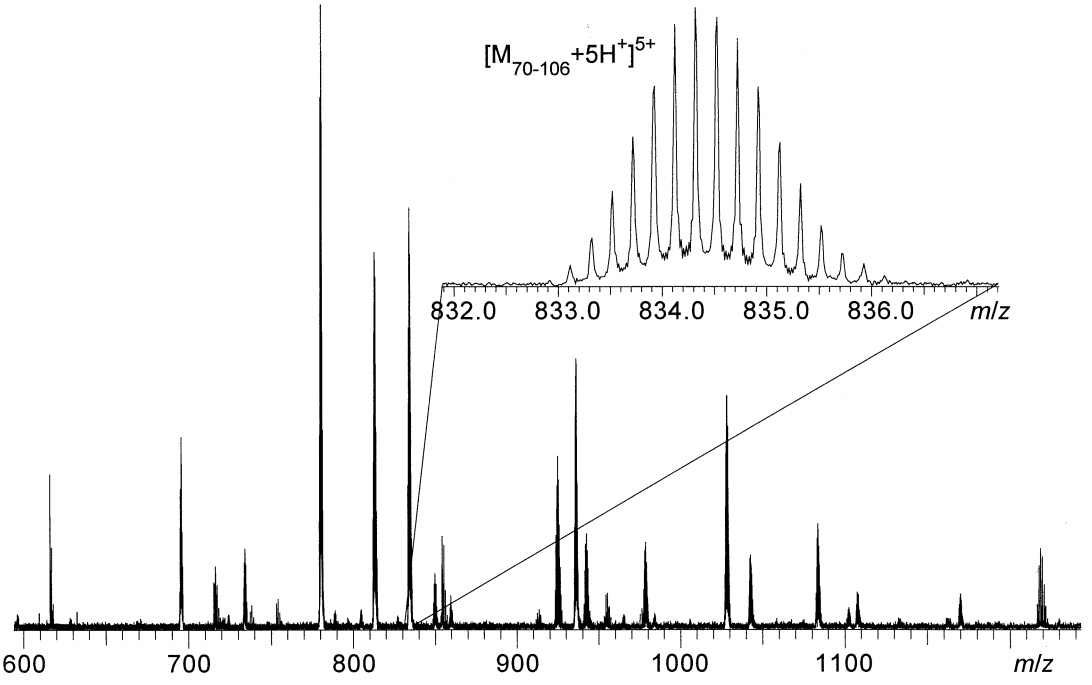

Figure 2. H/D exchange mass spectra of myoglobin peptic digest in 10% D_2O. The inset shows a set of isotopic peaks identified by AUTOHD as myoglobin fragment 70–106.

Results and Discussion

A typical H/D exchange FTICR mass spectrum of myoglobin, subjected to limited peptic digestion in 10% D_2O is shown in Figure 2. The inset shows a set of peaks identified by the AUTOHD program as an isotopic cluster of myoglobin peptic fragment 70–106 of charge 5+ with average amide deuterium content 0.104 (fraction of completely deuterated amide nitrogens). The corresponding output of the AUTOHD program is shown in Figure 3a and compared with original data in Figure 3b. The columns show the candidate (start and end residue numbers), the minimum χ^2 (eq 6) and the corresponding argmin(χ^2), the mass measurement error in ppm and the total χ^2, including the mass measurement and the corresponding argmin(χ^2). The significance of rejecting a candidate, often denoted 1-α, where α is the upper quantile of the χ^2-distribution, is reported in the last column in Figure 3. For an acceptable fit, this value is typically less than 0.95 and the χ^2-values are close to the number of observed isotopic peaks. What we can read from the numbers in Figure 3 is that the only reasonable peptide candidate is peptic fragment 70–106, given the binomial distribution of amide hydrogen/deuterium model and the in silico generated candidates. The low values on χ^2 and 1-α in this case, and others where the signal-to-noise ratio was high, indicates that the error in the measurements, or the constant c in eq 7, was slightly overestimated.

This identification agrees with what could be expected from an undeuterated control where this fragment was identified on accurate monoisotopic mass (data not shown). Commonly found peptic fragments include residues 1–11, 12–29, 30–69, 70–106, 107–137, and 138–153, covering 100% of the myoglobin sequence. All of these identifications agree with the undeuterated control, demonstrating the robustness of the approach.

Because of the relative lack of specificity of pepsin, the peptides often have to be subjected to sequencing by tandem mass spectrometry for identification [21]. However, the high mass accuracy of FTICR mass spectrometry allows masses to be determined within a few ppm, which significantly simplifies identification [50]. In most studies, peptic digests have been analyzed after chromatographic isolation of the individual peptides. During this step, a large amount of structural information is lost because of back exchange of deuterium to hydrogen, thus another advantage of using FTICR mass spectrometry is that many fragments can be detected simultaneously in the same mass spectrum. Even complex samples resulting from enzymatic digestion of large proteins can be analyzed without prior separation [41].

Side chains are known to have profound influence on the rate of exchange of amide hydrogens [35], and this has also been shown using mass spectrometry by Buijs et al. [36]. It was therefore interesting to apply the program to fragment mass spectra of smaller peptides. Substance P was chosen for this purpose, and Figure 4a shows a nozzle-skimmer CID spectrum of substance P after 2.1 min in 10% D_2O, after one day of incubation in 99.8% D_2O. The inset shows a typical identification of and fit to the b_9 ion in this spectrum. The average D content of the b_9 ion amounted to 0.688 while the amide hydrogens had $P(^2H)$ ranging from 0.33 to 0.85.

Electron capture dissociation is another fragmenta-

(a) *m/z* 832.882 $[M+5H^+]^{5+}$, M = 4164.408, 18 peaks found, max signal-to-background 167.6

fragment	minimum intensity χ^2	argmin χ^2	mass error (ppm)	minimum total χ^2	1-α
(62-99)	$1.45 \cdot 10^2$	0.076	-0.8	$1.45 \cdot 10^2$	1.000
(70-106)	$1.30 \cdot 10^1$	0.104	-8.6	$1.77 \cdot 10^1$	0.002
(10-46)	$6.16 \cdot 10^2$	0.315	-31.1	$6.76 \cdot 10^2$	1.000
(36-71)	$6.96 \cdot 10^2$	0.306	-33.9	$7.68 \cdot 10^2$	1.000
(47-84)	$1.42 \cdot 10^3$	0.349	-1.4	$1.41 \cdot 10^3$	1.000
(77-112)	$1.01 \cdot 10^4$	0.076	3.4	$1.01 \cdot 10^4$	1.000
(76-111)	$1.01 \cdot 10^4$	0.076	3.4	$1.01 \cdot 10^4$	1.000
(75-110)	$1.01 \cdot 10^4$	0.076	3.4	$1.01 \cdot 10^4$	1.000
(34-69)	$9.78 \cdot 10^3$	0.206	-16.6	$9.80 \cdot 10^3$	1.000
(22-57)	$9.78 \cdot 10^3$	0.239	-7.6	$9.78 \cdot 10^3$	1.000
(91-127)	$9.78 \cdot 10^3$	0.305	-6.5	$9.78 \cdot 10^3$	1.000
(14-49)	$9.78 \cdot 10^3$	0.349	-1.2	$9.78 \cdot 10^3$	1.000
(31-65)	$1.10 \cdot 10^5$	0.349	2.5	$1.10 \cdot 10^5$	1.000

(b)

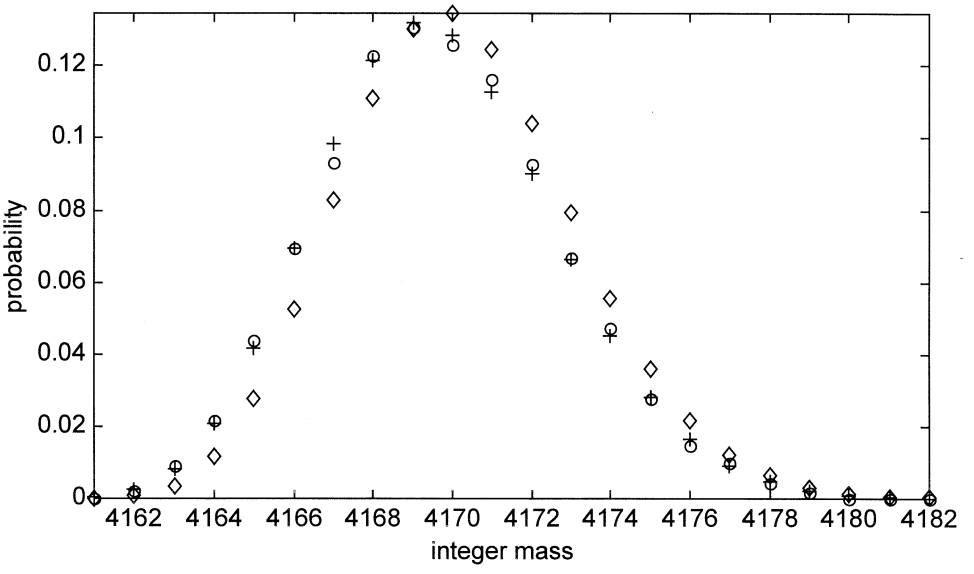

Figure 3. (**a**) Output from an AUTOHD program run on an H/D exchange mass spectra of myoglobin peptic digest and (**b**) graphical representation of the results. The columns in (**a**) display the candidate start and end residues, the χ^2 (from eq 6) and the corresponding argmin(χ^2), which corresponds to the average deuterium level, the mass measurement error in ppm, the total χ^2 including the mass measurement and the corresponding argmin(χ^2), and the significance of rejecting the candidate. In (**b**), circles are experimental values, crosses the best fit for fragment 70–106, diamonds best fit for fragment 62–99, the second best candidate.

tion technique of increasing importance in tandem mass spectrometry of peptides and proteins [32, 33]. Figure 4b shows an ECD spectrum of substance P after 7.9 min in 10% D_2O. The inset shows identification and fit to the c_{10} ion in this spectrum. The average D content of this c_{10} ion was found to be 0.295. The relatively poor fit to the experimental data, even given the low signal-to-noise ratio, indicates that the binomial model of amide H/D ratio is a poor approximation for this peptide. As could be expected, all b, y, and c fragments in these spectra were correctly identified by AUTOHD. For smaller peptides, such as substance P, the number of possible fragments is one or several orders of magnitude smaller than the number of possible peptic digest peptides of whole proteins, or 21 in the case of substance P versus 1239 for the myoglobin peptic digest. This means that the identification can often be done using measured accurate mass alone for small peptides. Larger proteins have more candidates near a given mass, posing an increased risk for multiple unrejectable candidates.

Analysis of mass spectra of peptides such as substance P takes fractions of a second, where as analysis of time series containing a few tens of spectra as complex as those of myoglobin digest, i.e., with a few hundred peaks per spectrum, takes on the order of tens of seconds on the computer systems used. As an example, complete analysis of peak lists from 14 myoglobin H/D exchange mass spectra containing in total 5766 peaks takes less than 8 s, using a vector size of 64 for the

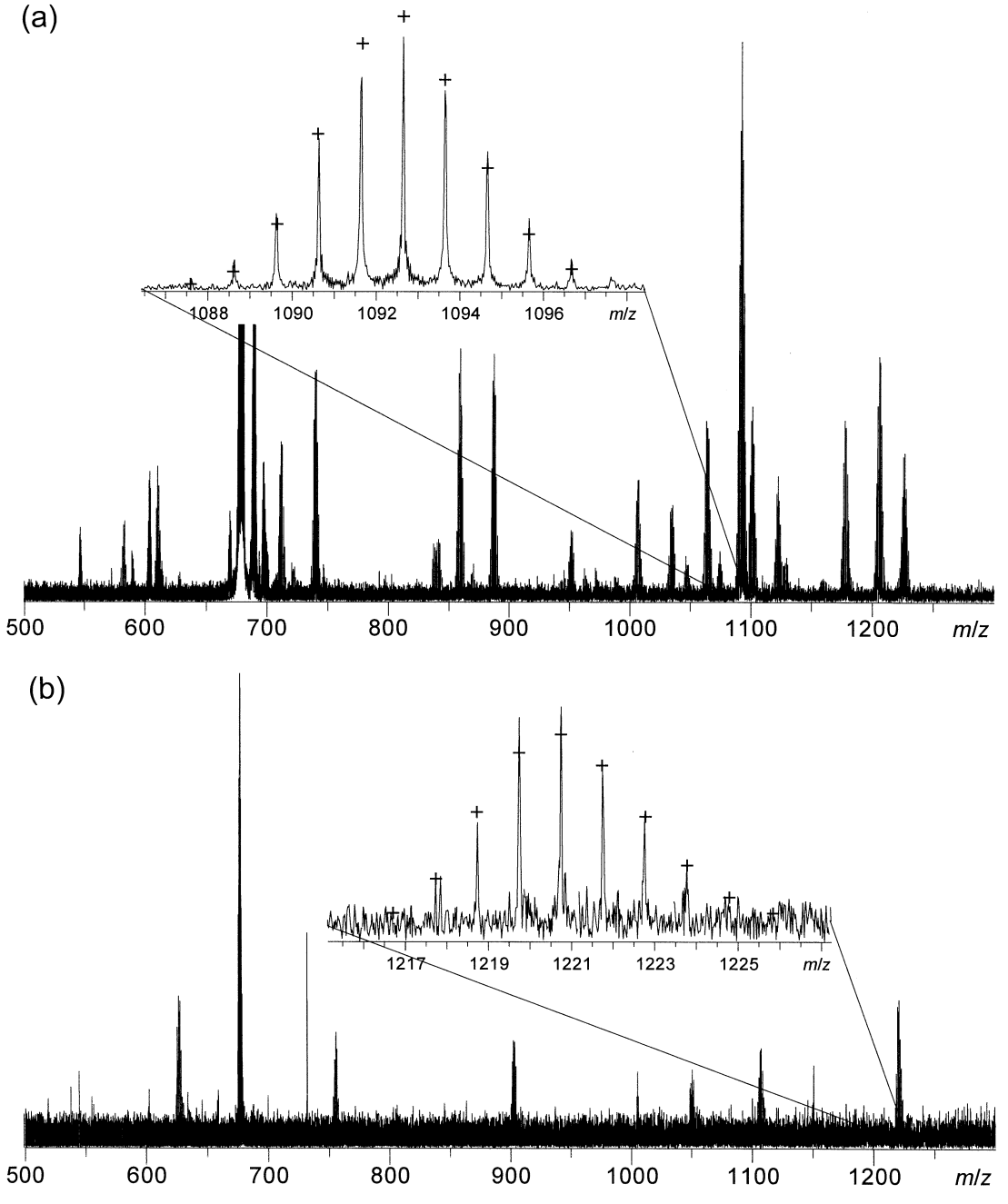

Figure 4. (**a**) Nozzle-skimmer CID of substance P after incubation 2.1 min in 10% D_2O. Inset shows the identification of b_+^9 with average amide deuterium content 0.688. (**b**) ECD of substance P after 7.9 min in 10% D_2O. Inset shows the identification of c_+^{10} with average amide deuterium content 0.295. Analysis of ECD spectra can be further complicated because of superposition of protonated and radical species [33].

calculations of isotopic distributions on a 1.1 GHz AMD Thunderbird PC system. The error due to folding over of the calculated isotopic distributions was less than 1 part in 10^7 in the relative intensities in the calculated distributions compared to using a vector size of 16,384, which is sufficiently large to hold all possible fragments without oversampling. This speed comes from using a fast FFT implementation, allowing oversampling of isotopic distributions and using a suitable minimization algorithm for the model of deuterium incorporation used. More parameters can be used in the model of deuterium incorporation, but the program would have to search in multiple dimensions. Such a multi-dimensional search or minimization requires more computation time than the one-dimensional model used here, but would still be feasible because of the efficiency of the program core.

To further demonstrate the applicability of the ap-

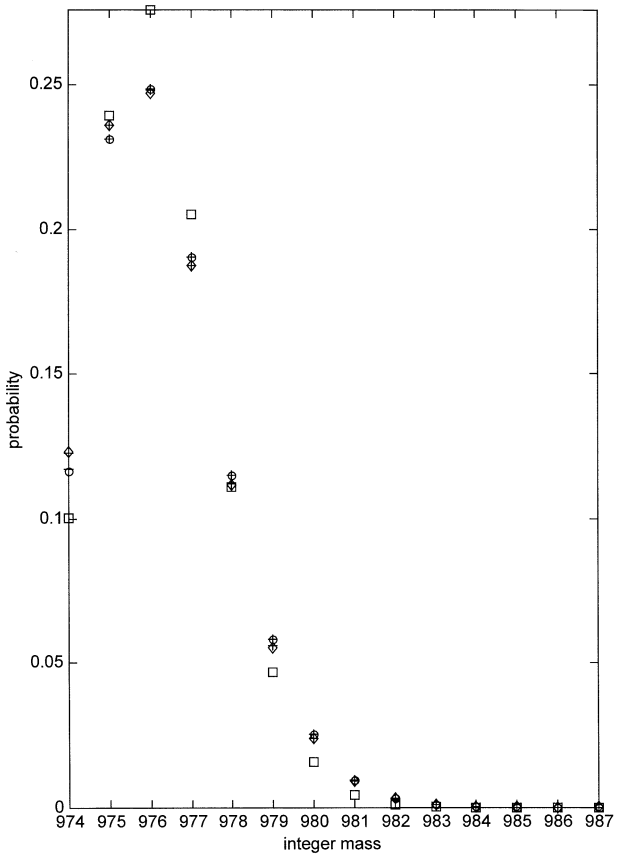

Figure 5. Fit by AUTOHD (crosses) to simulated spectra of a possible fragment (CKKSCCPCC) of human metallothionein with amide $P(^2H)$ chosen randomly and uniformly between 0.05 and 0.15. Mass spectra with (diamonds) and without (circles) error and noise were simulated. The average amide $P(^2H)$ was 0.102 (with error and noise) and 0.109 (without error and noise) respectively. The argmin(χ^2) was 0.103 and 0.109 respectively. The best fit to error and noiseless data using an averagine approximation of the elemental composition is shown as comparison (squares).

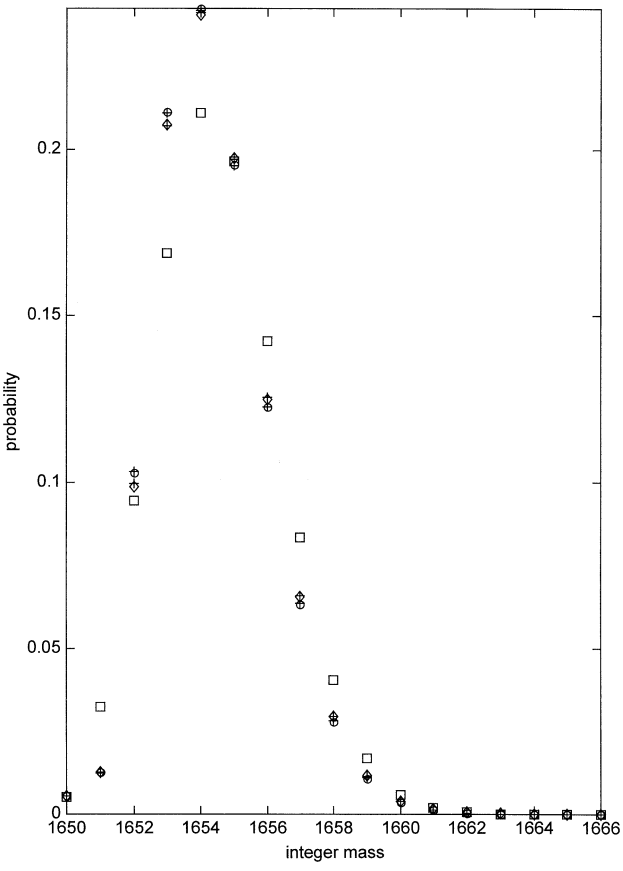

Figure 6. Simulation of fragment 12–20 (MKCSQCHTV) of human cytochrome *c* with the *c* heme covalently bound to the two cysteines. Amide $P(^2H)$ were chosen randomly and uniformly between 0.05 and 0.15. Mass spectra with (diamonds) and without (circles) error and noise were simulated. The average amide $P(^2H)$ was 0.098 (with error and noise) and 0.093 (without error and noise) respectively. The argmin(χ^2) was 0.098 and 0.093 respectively. The best fit to error and noiseless data using an averagine approximation of the elemental composition is shown as comparison (squares).

proach, a H/D exchange mass spectrum of the peptide CKKSCCPCC, which is one possible sulphur-rich fragment peptide of human metallothionein, was simulated by calculating the isotopic distribution using amide D/H ratios chosen at random uniformly between 0.05 and 0.15, with and without measurement errors and noise (Figure 5). The average amide $P(^2H)$ was 0.102 (with error and noise) and 0.109 (without error and noise) respectively. The argmin(χ^2) was 0.103 and 0.109 respectively. To show the importance of taking the elemental composition into account in these calculations, the averagine derived [9] elemental composition and isotopic distribution of peptides at this mass was used as a comparison. Using the correct number of labile and rapidly exchanging hydrogens and matching all the lowest masses, the argmin(χ^2) was 0.148 for the averagine comparison.

Similarly, Figure 6 shows the results of AUTOHD on a simulated isotopic distribution of fragment 12–20 (MKCSQCHTV) of human cytochrome *c* with the *c* heme covalently bound to the two cysteines. A spectrum of enzymatic digest fragments of cytochrome *c* with heme Fe^{III} has been reported previously [41]. The simulations were performed as described above for the metallothionein fragment. The average $P(^2H)$ used was 0.098 (with error and noise) and 0.093 (without error and noise). The argmin(χ^2) found was 0.098 and 0.093 respectively. The argmin(χ^2) for the averagine fit was 0.377 (constrained by the matching of all the lowest masses).

The averagine model was used as comparison only to emphasize the effect of taking the exact elemental composition into account—this model was never intended to be used for cases where the elemental composition was known or to extract more information than the monoisotopic mass from measured isotopic distributions [9] and these results do not reflect poorly upon either the averagine model itself or methods using it [9, 24].

To estimate the applicability of this approach in the

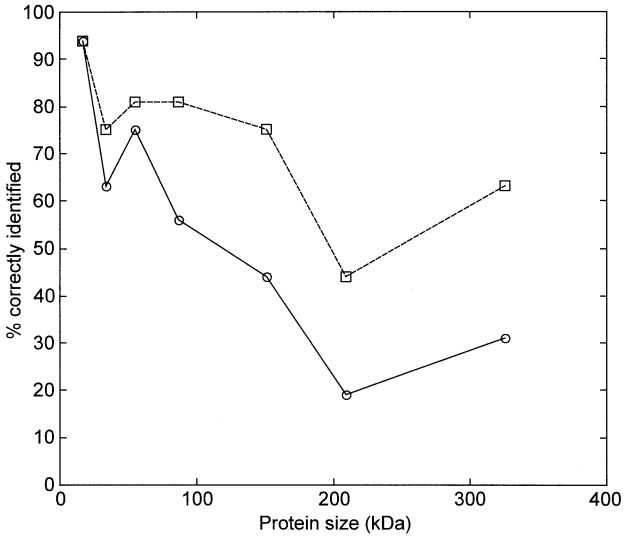

Figure 7. Simulation of the program and single- $P(^2H)$ model on a number of hypothetical fusion proteins with myoglobin. The squares are the fraction of myoglobin fragments that would be correctly identified and the circles the fraction that would be correctly and unambiguously identified in such fusion proteins. Data from three different myoglobin mass spectra were used in this simulation.

analysis of enzymatic digests of larger proteins, hypothetical fusion proteins of myoglobin and a number of other proteins, chosen according to size but otherwise at random, from a small protein database were simulated in the computer. These proteins were lysozyme (PID g4557894), α_2-HS-glycoprotein (g2116653), serum albumin (g4502027), epidermal growth factor (g4503491), complement factor C4 (g116602) and the von Willebrand factor (g4507907), all human. The sizes of the resulting fusion proteins were 54, 86, 151, 209, and 326 kDa respectively. These fusion products were then used as input to DIG, and the three mass spectra of myoglobin after 0.1, 120, and 2880 min in D_2O were used to evaluate the performance of AUTOHD. The program output for the six commonly observed fragments cited above was observed. The maximum signal-to-background ratio of these fragments varied from around 3 to 250. The result of this simulation is shown in Figure 7 where the squares are the fraction of correctly identified and the circles the fraction of correctly and unambiguously identified fragments. Although the program made more erroneous assignments in the larger proteins, the simulation indicates that this program and model is also applicable in the analysis of proteins significantly larger than 15 kDa. The mass spectra used in this simulation were not recorded and calibrated in an optimal way, resulting in mass measurement errors around 5–10 ppm. Obtaining a higher mass accuracy should significantly improve the performance of the program, and mass accuracies down to 1.6 ppm have been reported previously for complex enzymatic digests using this instrument [41].

Conclusions

The benefit of this program in the presented version is twofold. First, it rapidly and automatically identifies and calculates H/D ratios of species in H/D exchange mass spectra of proteins and peptides, which reduces the amount of manual labor in the analyses of such spectra. Second, it makes the identification independent on undeuterated controls and is able to resolve ambiguities in identification based on accurate mass alone. This is important, since reliability of such controls depends on keeping the experimental conditions for incubation, quenching, digesting, mixing, and mass spectrometry constant, so that the same fragments are observed every time. In our experience this is not always the case; for example the intensity of the most C-terminal peptic fragment of myoglobin varied considerably from experiment to experiment and was sometimes difficult to detect. The controls are used to calculate the deuterium incorporation from measured average masses.

The approach described in this paper could be applied in high-throughput screens of protein stability [51] or adopted for automatic identification and characterization of components in complex mass spectra from other isotopic labeling experiments of proteins [52, 53]. Future versions of the program will include input of an model of arbitrary isotope incorporation, including maximum entropy methods applied on distributions of H/D exchange rate constants [34], and require only replacement of eq 4 and the minimization algorithm. In particular, using models with more parameters is one way to achieve information on the distribution of exchange rates within a peptic fragment. In the case of multiple distributions for the same fragment, for instance bimodal distributions [19], the program in its present form would probably not find a significant match. Instead, a model consisting of two (or more) independent distributions, each with its own set of parameters, and the relative probabilities of the peptide amide hydrogens having any one of these distributions could be used. The program with the binomial model separates those peptic fragments that can be approximated with a binomial distribution from those that cannot. The computer programs are available from the authors.

Acknowledgments

The authors gratefully acknowledge the support from the Swedish Natural Sciences Research Council (Grant K-1618/1999) and the Knut and Alice Wallenberg Foundation. The authors also thank Lars-Larsson Cohn at the Department of Mathematics and Henrik Brandén at the Department of Information Technology, both at Uppsala University, for valuable discussions, and Mikael Danfelter, Margareta Ramström, and Youri Tsybin, all at the Division of Ion Physics, for kindly contributing experimental data.

References

1. Pandey, A.; Mann, M. *Nature* **2000,** *405,* 837–846.
2. McLafferty, F. W.; Fridriksson, E. K.; Horn, D. M.; Zubarev, R. A.; Lewis, M. A. *Science* **1999,** *284,* 1289–1290.
3. Rostom, A. A.; Robinson, C. V. *Curr. Opin. Struct. Biol.* **1999,** *9,* 135–141.
4. Henderson, S. C.; Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *Anal. Chem.* **1999,** *71,* 291–301.
5. Last, A. M.; Robinson, C. V. *Curr. Opin. Chem. Biol.* **1999,** *3,* 564–570.
6. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998,** *17,* 1–35.
7. Conrads, T. P.; Anderson, G. A.; Veenstra, T. D.; Pasa-Tolic, L.; Smith, R. D. *Anal. Chem.* **2000,** *72,* 3349–3354.
8. Yergey, J.; Heller, H.; Hansen, G.; Cotter, R. J.; Fenselau, C. *Anal. Chem.* **1983,** *55,* 353–356.
9. Senko, M. W.; Beu, S. C.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 229–233.
10. He, F.; Hendrickson, C. L.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **2000,** *11,* 120–126.
11. Shi, S. D. H.; Hendrickson, C. L.; Marshall, A. G. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95,* 11532–11537.
12. Englander, S. W.; Sosnick, T. R.; Englander, J. J.; Mayne, L. *Curr. Opin. Struct. Biol.* **1996,** *6,* 18–23.
13. Smith, D. L.; Deng, Y.; Zhang, Z. *J. Mass Spec.* **1997,** *32,* 135–146.
14. Woodward, C. *J. Am. Soc. Mass Spectrom.* **1999,** *10,* 672–674.
15. Roder, H.; Elove, G. A.; Englander, S. W. *Nature* **1988,** *335,* 700–704.
16. Miranker, A.; Robinson, C. V.; Radford, S. E.; Aplin, R. T.; Dobson, C. M. *Science* **1993,** *262,* 896–900.
17. Katta, V.; Chait, B. T. *J. Am. Chem. Soc.* **1993,** *115,* 6317–6321.
18. Dharmasiri, K.; Smith, D. L. *J. Am. Soc. Mass Spectrom.* **1997,** *8,* 1039–1045.
19. Zhang, Z.; Post, C. B.; Smith, D. L. *Biochemistry* **1996,** *35,* 779–791.
20. Johnson, R. S. *J. Am. Soc. Mass. Spectrom.* **1996,** *7,* 515–521.
21. Maier, C. S.; Kim, O. H.; Deinzer, M. L. *Anal. Biochem.* **1997,** *252,* 127–135.
22. Mandell, J. G.; Falick, A. M.; Komives, E. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95,* 14705–14710.
23. Chung, E. W.; Nettleton, E. J.; Morgan, C. J.; Gross, M.; Miranker, A.; Radford, S. E.; Dobson, C. M.; Robinson, C. V. *Protein Sci.* **1997,** *6,* 1316–1324.
24. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass. Spectrom.* **2000,** *11,* 320–332.
25. Zhang, Z.; Smith, D. L. *Protein Sci.* **1993,** *2,* 522–531.
26. Deng, Y.; Smith, D. L. *J. Mol. Biol.* **1999,** *294,* 247–258.
27. Waring, A. J.; Mobley, P. W.; Gordon, L. M.; *Proteins* 1998, *Suppl.,* 38–49.
28. Anderegg, R. J.; Wagner, D. S.; Stevenson, C. L.; Borchardt, R. T. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 425–433.
29. Hayes, R. N.; Gross, M. L. *Methods Enzymol.* **1990,** *193,* 237–263.
30. Cody, R. B.; Freiser, B. S. *Int. J. Mass Spectrom. Ion Phys.* **1982,** *41,* 199–204.
31. Deng, Y.; Pan, H.; Smith, D. L. *J. Am. Chem. Soc.* **1999,** *121,* 1966–1967.
32. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998,** *120,* 3265–3266.
33. Axelsson, J.; Palmblad, M.; Håkansson, K.; Håkansson, P. *Rapid Commun. Mass Spectrom.* **1999,** *13,* 474–477.
34. Zhang, Z.; Li, W.; Logan, T. M.; Li, M.; Marshall, A. G. *Protein Sci.* **1997,** *6,* 2203–2217.
35. Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W. *Proteins* **1993,** *17,* 75–86.
36. Buijs, J.; Håkansson, K.; Hagman, C.; Håkansson, P.; Oscarsson, S. *Rapid Commun. Mass Spectrom.* **2000,** *14,* 1751–1756.
37. Palmblad, M.; Håkansson, K.; Håkansson, P.; Feng, X.; Cooper, H. J.; Giannakopulos, A. E.; Green, P. S.; Derrick, P. J. *Eur. Mass. Spectrom.* **2000,** *6,* 267–275.
38. Rockwood, A. L. *Rapid Commun. Mass Spectrom.* **1995,** *9,* 103–105.
39. Cooley, J. W.; Tukey, J. W. *Math. Comp.* **1965,** *19,* 297–301.
40. http://prospector.ucsf.edu/ *ProteinProspector Home Page.*
41. Palmblad, M.; Wetterhall, M.; Markides, M.; Håkansson, P.; Bergquist, J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1029–1034.
42. Suckau, D.; Shi, Y.; Beu, S. C.; Senko, M. W.; Quinn, J. P.; Wampler, F. M. D.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90,* 790–793.
43. Chowdhury, S. K.; Katta, V.; Chait, B. T. *J. Am. Chem. Soc.* **1990,** *112,* 9012.
44. Brent, R. P. *Algorithms for Minimization without Derivatives*; Prentice-Hall: New York, 1973; p 78.
45. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C, the Art of Scientific Computing, 2nd ed.*; Cambridge, UK: Cambridge University Press, pp 402–405.
46. Frigo, M.; Johnson, S. G. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. FFTW: An Adaptive Software Architecture for the FFT Vol. 3; Piscataway, NJ: IEEE, 1998 pp 1381–1384.
47. http://caml.inria.fr *The Caml Language Home Page.*
48. http://gcc.gnu.org *GCC Home Page.*
49. http://www.cygwin.com *Cygwin Home Page.*
50. Zubarev, R. A.; Håkansson, P.; Sundqvist, B. U. R. *Anal. Chem.* **1996,** *68,* 4060–4063.
51. Ghaemmaghami, S.; Fitzgerald, M. C.; Oas, T. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97,* 8296–8301.
52. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nature Biotech.* **1999,** *17,* 994–999.
53. Geng, M.; Junyan, J.; Regnier, F. E. *J. Chromatogr. A.* **2000,** *870,* 295–313.